THE EFFECT OF HEAD-RELATED FILTERING AND EAR-SPECIFIC DECODING BIAS ON AUDITORY ATTENTION DETECTION

Authors

Neetha Das (Corresponding Author)

Dept. Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Dept. Neurosciences, ExpORL, KU Leuven, Herestraat 49 bus 721, B-3000 Leuven, Belgium

email: neetha.das@esat.kuleuven.be

Wouter Biesmans

Dept. Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Alexander Bertrand

Dept. Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Tom Francart

Dept. Neurosciences, ExpORL, KU Leuven, Herestraat 49 bus 721, B-3000 Leuven, Belgium

September 15, 2016

The effect of head-related filtering and ear-specific decoding bias on auditory attention detection

Neetha Das[†⋆], Wouter Biesmans[†], Alexander Bertrand[†], Tom Francart[⋆]

Abstract

Objective. We consider the problem of auditory attention detection (AAD), where the goal is to detect which speaker a person is attending to, in a multi-speaker environment, based on neural activity. This work aims to analyze the influence of head-related filtering and ear-specific decoding on the performance of an AAD algorithm. *Approach.* We recorded high-density EEG of 16 normal-hearing subjects as they listened to two speech streams while tasked to attend to the speaker in either their left or right ear. The attended ear was switched between trials. The speech stimuli were administered either dichotically, or after filtering using head-related transfer functions (HRTFs). A spatio-temporal decoder was trained and used to reconstruct the attended stimulus envelope, and the correlations between the reconstructed and the original stimulus envelopes were used to perform AAD, and arrive at a percentage correct score over all trials. *Main results.* We found that the HRTF condition resulted in significantly higher AAD performance than the dichotic condition. However, speech intelligibility, measured under the same set of conditions, was lower for the HRTF filtered stimuli. We also found that decoders trained and tested for a specific attended ear performed better, compared to decoders trained and tested for both left and right attended ear simultaneously. In the context of the decoders supporting hearing prostheses, the former approach is less realistic, and studies in which each subject always had to attend to the same ear may find over-optimistic results. *Significance.* This work shows the importance of using realistic binaural listening conditions and training on a balanced set of experimental conditions to obtain results that are more representative for the true AAD performance in practical applications.

Index Terms

Auditory attention detection, EEG processing, neuro-steered auditory prostheses, brain-computer interface, cocktail party, speech stimuli, stimulus reconstruction, acoustic conditions.

I. Introduction

In a multi-speaker environment, a normal-hearing person can focus his attention towards one of the speakers, while ignoring other speakers. The corresponding neural mechanism is an important topic of research, not only due to the fact that it will bring us one step closer to understanding the human brain, but also because of the potential it holds in the realization of neural feedback to support auditory prostheses. More specifically, the task

This research work was carried out at the ESAT and ExpORL Laboratories of KU Leuven, in the frame of KU Leuven Special Research Fund BOF/STG-14-005 and OT/14/119. The work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 637424). The scientific responsibility is assumed by its authors.

[†] KU Leuven, Dept. Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics. Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.

[⋆] KU Leuven, Dept. Neurosciences, ExpORL. Herestraat 49 bus 721, B-3000 Leuven, Belgium.

of detecting which speaker a person is attending to is termed auditory attention detection (AAD). It has been successfully demonstrated based on electroencephalography (EEG) and magnetoencephalography (MEG) in a two-speaker scenario [1]–[4]. The technology to decode auditory attention unobtrusively could pave the way towards neuro-steered hearing devices, where AAD information can be used to steer a noise suppression algorithm towards the attended speaker [5]. Beginning with scientifically establishing the crucial role that the envelope of speech plays in speech recognition [6], [7], and later studies [8]–[11] showing that cortical oscillations phase-lock to the speech envelope, research has progressed significantly in deciphering the auditory attention problem. Recent developments have addressed the cocktail party scenario, where cortical responses are seen to track the attended speech envelope better than the unattended speech envelope [3], [12], [13].

AAD can be performed by extracting relevant features from the EEG recordings and using them to solve a classification problem [14]. Another approach is to use trained spatio-temporal linear decoders to reconstruct the envelopes of the attended speech stream from the EEG recordings. The reconstructed stimulus is then correlated with the actual stimulus speech envelope. The 'attended to' speaker is chosen as the one corresponding to the speech envelope with the higher correlation with the reconstructed envelope. The decoders are trained using EEG recordings as well as the stimulus envelopes based on a least-squares (LS) regression [2], or by maximizing a cross correlation ratio [3].

Although many research groups actively study AAD, there is little uniformity in experimental protocols and acoustic conditions under which the EEG recordings are obtained. In some studies, the stimuli are presented dichotically, i.e., one separate speech stream to each ear [2], [15], or are mixed into a single acoustic stream which is played at both ears [3]. In some other studies they are administered through loudspeakers placed at specific positions in the room [14], or interaural level differences are applied to the stimuli to mimic those of sound sources at specific positions with respect to the speaker [4]. Kerlin et al. [12] used stimuli filtered using head-related transfer functions (HRTFs) to simulate sources at different positions in the room. In addition to these differences in the presentation of the stimuli, different research groups use different protocols to instruct the subject to which speaker they should attend. Some studies use neural recordings during which a subject had to focus on only one ear/speaker throughout the experiment [2], [4], [15], while in other studies, the subject had to switch focus between ears/speakers [12], [14]. So far, it is unclear whether, and how much, the type of stimulus presentation and/or the use of only a single attended ear in experimental protocols has an impact on the AAD performance, or whether they introduce bias in the results. In this paper, we aim to investigate whether such bias appears, and, if so, quantify the significance.

We follow the approach of stimulus reconstruction based on least-squares error minimization [2]. For the experiment, stimuli are administered to the subject under two conditions: dichotic - where the two unfiltered speech streams are administered separately to each ear, and HRTF-filtered, where the stimuli are filtered using a head-related transfer function to simulate the position of speakers at 90 degrees to the left and right of the subject. Each subject is instructed to alternately focus towards his/her left or right side, for each of the presentations in the experiment. We analyze the performance of AAD, when decoders are trained and tested using cortical EEG recordings. The

analysis is done on subject-specific (SS) decoders as well as on the more general subject-independent/ generic decoders.

The paper is organized as follows. In Section II, we describe the experimental details, how the acquired data is processed, decoder design and the different tested conditions. In Section III, we provide the AAD performance based on differences in stimuli, and ear-specific decoding. The implications of these findings are discussed in Section IV. Finally, we summarize and draw conclusions in Section V.

II. Methods

A. Participants

Eight male and eight female normal-hearing volunteers, between 17 and 30 years of age, participated in the experiment. All subjects had to fill out a modified version of a questionnaire [16] to assess their handedness and ear preference. The responses to the questionnaire showed that all subjects were right-handed. While three subjects showed ambilateral ear preference, and one subject showed a left ear preference, all other subjects showed a right ear preference. Normal hearing for all subjects was verified by pure-tone audiometry. Every subject signed an informed consent form approved by the local ethical committee.

B. Data Acquisition

EEG recordings were made in a soundproof, electromagnetically shielded room. The BioSemi ActiveTwo system was used to record 64-channel EEG signals at 8196 Hz sample rate. The audio signals, low pass filtered at 4 kHz, were administered to each subject at 60 dBA through a pair of insert phones (Etymotic ER3A). The experiments were conducted using the APEX 3 program developed at ExpORL [17].

C. Stimuli and Procedures

Four Dutch short stories [18], narrated by different male speakers, were used as stimuli. All silences longer than 500 ms in the audio files were truncated to 500 ms. Each story was divided into two parts of approximately 6 minutes each. During a presentation, the subjects were presented with the six-minutes part of two (out of four) stories played simultaneously. There were two stimulus conditions, i.e., ‘HRTF’ or ‘dichotic’ (see below). An experiment here is defined as a sequence of 4 presentations, 2 for each stimulus condition and ear of stimulation, with questions asked to the subject after each presentation. All subjects sat through three experiments within a single recording session. An example for the design of an experiment is shown in Table I. The first two experiments included four presentations each. During a presentation, the subjects were instructed to listen to the story in one ear, while ignoring the story in the other ear. After each presentation, the subjects were presented with a set of multiple-choice questions about the story they were listening to in order to help them stay motivated to focus on the task. In the next presentation, the subjects were presented with the next part of the two stories. This time they were instructed to attend to their other ear. In this manner, one experiment involved four presentations in which the subjects listened to a total of two stories, switching attended ear between presentations. The second experiment had the same design

but with two other stories. Note that the Table I was different for each subject or recording session, i.e., each of the elements in this table were permuted between different recording sessions to ensure that the different conditions (stimulus condition and the attended ear) were equally distributed over the four presentations. Finally, the third experiment included a set of presentations where the first two minutes of the story parts from the first experiment, i.e. a total of four shorter presentations, were repeated three times, to build a set of recordings of repetitions. Thus, a total of 72 minutes of EEG was recorded per subject. Please note that these repetitions were included for a specific purpose in a related study. In this work, these data were not treated as repetitions, and all analysis was carried out on a single-trial basis even though a subset of stimuli appear multiple times in the dataset.

Presentation	Left Stimulus	Right Stimulus	Attended Ear	Stimulus Condition
1	Story1, part1	Story2, part1	Left	Dichotic
2	Story2, part2	Story1, part2	Right	HRTF
3	Story2, part1	Story1, part1	Left	Dichotic
4	Story1, part2	Story2, part2	Right	HRTF

TABLE I: An example of the design of experiments 1 and 2

Throughout the experiments, the attended ear of the subject was switched between presentations to obtain equal amounts of data for both left and right attended ear per subject. Thus, approximately 36 minutes of EEG recording per attended ear were obtained for each subject. Each presentation also had its unique stimulus condition. These will be referred to as 'dichotic' and 'HRTF' conditions. The 'dichotic' condition is where the audio streams were administered, at equal intensities, to separate channels of the insert phones without any filtering. The 'HRTF' condition is where the two audio streams were filtered by head-related transfer functions, simulating an auditory environment where each speaker was perceived to be located 90 degrees to the left and right of the subject. This is a more realistic scenario, where stimuli to each ear contain both audio streams. The HRTFs, used to filter the stimuli, were measured on a dummy head in an anechoic room using in the ear (ITE) microphones. These stimulus conditions were also balanced over different sessions. The order of conditions, and attended stories were randomized across subjects. All stimuli were normalized to have the same root-mean-square value.

D. Data Preprocessing

The EEG signal was filtered with an equiripple bandpass filter with passband attenuation of 0.5 dB and stopband attenuation of 20 and 15 dB. The filter's passband was between 1 and 9 Hz, which is the frequency range of most interest for cortical tracking of speech stimuli [1], [11], [19]. It is noted that several earlier studies on EEG-based AAD excluded the 1-2 Hz band in the analysis [2], [4], [20], [21]. However, we observed that including the 1-2 Hz band indeed significantly improved the AAD performance in this study, and found the best performance for a

passband between 1-9 Hz. After bandpass filtering, the signal was downsampled to 20 Hz. To extract envelopes from the speech stimuli, we relied on an auditory filterbank with power law compression, as proposed in [20], [21]. To this end, the speech signal was first fed to a gammatone filterbank [22], [23] after which a power law compression with exponent 0.6 was applied on the output signal of each subband. For the output signal $x_k(t)$ of subband k at time t, this means we computed $|x_k(t)|^{0.6}$. Each subband was then bandpass filtered with the same filter as used in the EEG recordings, and downsampled to 20 Hz. The resulting subband envelopes were then summed to construct a single envelope [21].

The recorded data set was divided into 30 second trials. Thus we analyzed 152 trials per subject, balanced over the two stimulus conditions, and attended ears, as explained in section II-C.

E. Stimulus Reconstruction and Decoder Design

1) Stimulus Reconstruction: The basic framework for stimulus reconstruction is similar to the algorithm described in [2]. The envelopes extracted from the attended and unattended speech streams are denoted by $s_a(t)$ and $s_u(t)$ respectively. We attempt to reconstruct the attended speech stream using a linear spatio-temporal decoder applied to the EEG data. If $D(n, c)$ denotes the decoder weight for channel c at time lag index n, and $M(t, c)$ denotes the EEG recording from channel c at time index t, the attended speech envelope is reconstructed from N_l time lags and C channels of EEG data as:

$$\tilde{s}_a(t) = \sum_{n=0}^{N_l-1} \sum_{c=1}^{C} D(n, c)M(t + n, c). \tag{1}$$

The time lag represents the difference between the actual time of administering the speech stimulus, and the time when the cortical activity reflects the dynamics of the speech stimulus. Here, time lags of up to 250 ms are used for an effective reconstruction of the speech envelope [2].

2) Decoder Design: The decoder is designed in such a way as to minimize the mean squared error between the original and reconstructed attended speech envelopes over a training set, consisting of a subset of the trials. This leads to the analytical solution of the standard minimum mean squared error problem:

$$D = R_{MM}^{-1}\mathbf{r}_{MS}, \tag{2}$$

where R_{MM} is the spatio-temporal correlation matrix of the EEG data over all channels and time lags, and R_{MS} is the cross-correlation vector between the attended speech envelope and the EEG data, over all channels and time lags.

Every 30-second trial results in one R_{MM} and $\mathbf{r}_{MS}$ pair. These matrices are averaged over all the trials that are used for training the decoder. It is finally the averaged correlation matrices and the average cross-correlation vectors that are used in equation (2) to compute the decoder. This is different from the standard approach [2] where equation (2) is computed for every trial independently, after which the decoders from all trials are averaged. The current approach leads to a correlation matrix that is better conditioned than the per-trial correlation matrices. Furthermore,

if sufficient training data is available, this eliminates the need to apply regularization and hence the tuning of a regularization parameter [21]. Averaging the correlation matrices corresponds to optimizing a single decoder to minimize the mean square error over the entire training data set, rather than averaging trial-specific decoders that minimize the mean square error per trial [21]. We build decoders based on two approaches - 'subject specific' where a decoder is trained based on all the trials from the same subject, except for the single trial under test (leave-one-trial-out); and 'generic' where the decoder is trained on every other trial from every other subject (leave-one-subject-out). Pearson's correlation coefficients are then computed for the test trial between the reconstructed envelope $\tilde{s}_a$ and, s_a and s_u. If the attended stimulus envelope has a higher correlation with the reconstructed envelope, compared to the unattended stimulus envelope, the decoding is assumed to be a success. Decoding accuracy is the percentage of correctly decoded trials across all the trials for each subject.

III. Results

We trained decoders and tested the performance of AAD under five different analysis conditions listed below. Decoders were tested on 30 second trials, with a total of 36 minutes of data per subject per condition.

1) 'Dichotic' condition: Only trials with dichotic stimuli are used for training and testing, irrespective of the attended ear, i.e., decoders are trained over both right-attended and left-attended trials.
2) 'HRTF' condition: The same as above, but now, only trials with HRTF-filtered stimuli are used.
3) 'Same ear' condition: Only trials in which the subject attended to the same ear as the test trial were used for training. No distinction is made between dichotic or HRTF conditions, i.e., trials from both conditions are used for training.
4) 'Both ears' condition: The training set contains trials from all conditions (left attended, right attended, dichotic and HRTF).
5) 'Other ear' condition: Only trials in which the subject attended to the other/opposite ear as the test trial were used for training. Therefore, decoders trained on left-attended trials are tested on right-attended trials, and vise versa. No distinction is made between dichotic or HRTF conditions, i.e., trials from both conditions are used for training.

A. Effect of attended ear conditions

According to the Wilcoxon signed rank test, the 'same ear' trained decoders performed significantly better than 'both ears' trained decoders for both generic ($p < 0.001$) and subject specific ($p < 0.001$) approaches. As can be seen in Fig. 1, the 'same ear' condition resulted in an increase in median decoding accuracy of 6.9% for generic and 2.3% for subject specific decoders, in comparison to the 'both ears' condition. Also, the 'same ear' decoders trained and tested on right ear trials performed better than those based on left ear trials ($p = 0.006$ by Wilcoxon's signed rank test) by 5.2% for the generic decoding approach. No significant difference between the left and right ear decoders was found when using subject-specific decoders. Furthermore, we observed no statistical differences between the AAD performance when decoding right ear trials using left ear decoders, and decoding left ear trials

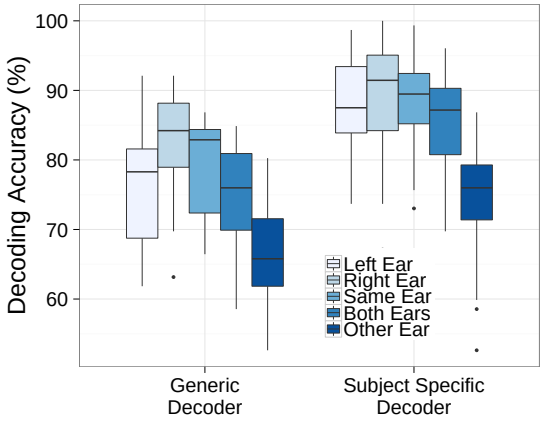

Fig. 1: Decoding accuracies of 16 subjects under different attended ear conditions: 'Left Ear' or 'Right Ear' - decoders trained and tested on trials where, per subject, attended ear is left only or right only, respectively; 'Same Ear' - results averaged over both 'Left Ear' and 'Right Ear' conditions; 'Both Ears' - decoders trained and tested for all trials; 'Other Ear' - decoder trained on trials attending to the opposite ear, from the test trial.

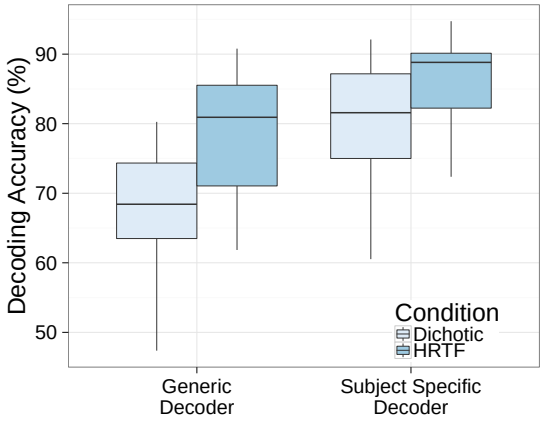

Fig. 2: Decoding accuracies of 16 subjects under dichotic and HRTF stimulus conditions, using generic and subject specific decoders.

using right ear decoders. We therefore combined these two results in the 'other ear' condition. In comparison to the 'both ears' condition, the 'other ear' condition resulted in 10.2% poorer performance for the generic decoding ($p < 0.001$) approach, and 11.2% poorer performance for the subject specific ($p < 0.001$) approach.

B. Effect of different stimulus conditions

The performance of the different stimulus conditions (dichotic vs. HRTF) can be seen in Fig. 2. According to the Wilcoxon signed rank test, significantly better performance was achieved for the HRTF condition for both generic ($p < 0.001$) and subject specific ($p = 0.002$) approaches. With subject specific decoders, the HRTF condition resulted in

an increase in median accuracy of 7.2%, with respect to the dichotic condition. With generic decoders, the increase was 12.5%. Thus, in both approaches, AAD under the HRTF condition showed nearly 10% better performance compared to the dichotic condition.

IV. Discussion

With a median decoding accuracy of 76.0% using generic decoders, and 87.2% using subject specific decoders, the auditory attention detection system developed based on the experiments described in Section II achieved performance comparable to the literature [2]–[4]. We compared the performance of the algorithm under different attended ear and stimulus conditions.

A. Effect of attended ear and ear-specific decoding

It is seen that decoders trained and tested under the 'same ear' condition perform better than those under 'both ear' condition. Working towards the aim of decoders that may one day support noise suppression in hearing devices, a trained decoder must be able to effectively decode attention irrespective of which speaker the subject is focusing on. In studies that use only one attended ear per subject during their data collection phase, the subject specific decoders are trained only under the 'same ear' condition, rather than the required 'both ear' condition. Thus, these decoders can lead to over-optimistic results. Such a decoder, when faced with the EEG recordings of the subject attending to the ear that it is not trained to recognize ('other ear' condition), will tend to perform worse than a decoder that is trained on both ears. On the other hand, 'both ear' decoders are more generic as they are trained on both left-ear and right-ear attended trials, and therefore give a more realistic estimate of the performance. Many studies rely on experiments where half the subjects are instructed to listen to the right speaker, and half to the left speaker. Our study demonstrates that the algorithms designed on such data may be expected to have a ear-specific decoding bias, resulting in a subject-specific decoding performance that is higher than reality.

It is also observed that decoders trained and tested on right-ear attended trials resulted in better AAD performance compared to those trained and tested on left-ear attended trials. It could be due to a well-known phenomenon termed the right ear advantage [24], which can be observed in dichotic listening tests. It can be observed when different speech stimuli are presented to the two ears of a subject, and more words arriving at the right ear are correctly reported than those arriving at the left ear. It is often associated with the crossed pathways of the ear to the auditory cortex dominating the uncrossed, and the left hemisphere being critical for speech recognition [25].

In order to investigate the laterality of neural responses correlated to the attended stimulus, we analyzed the temporal response functions (TRFs) (obtained using the AESPA technique in [10]) at different channels trained under different conditions. The TRF coefficients for EEG channel k are obtained from a linear least-squares fit between multiple time-lagged versions of the stimulus on the one hand, and the recorded EEG signal at channel k on the other hand. We computed TRFs for every 6.5 minutes of continuous recording, and averaged the coefficients over all subjects (within conditions) . The TRF coefficients at each lag and each channel can be visualized using multiple topoplots, i.e., one for each time lag. As can be seen in Fig. 3, during lags 160-200 ms, which are also

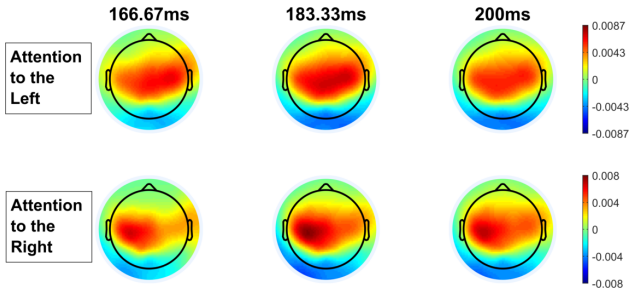

Fig. 3: TRF coefficients at different lags.

the lags at which attention detection performance peaks, we observed activity spread over the two hemispheres in the 'attended left' case, whereas the activity was more concentrated towards the left hemisphere in the 'attended right' case, similar to what can be seen in [15]. We believe this could also be attributed to the left hemisphere's dominance in speech processing [26], together with contralateral auditory pathways yielding larger responses than ipsilateral auditory pathways.

B. Effect of head-related filtering

AAD accuracy is found to be higher when the stimuli are HRTF-filtered, compared to the dichotic condition. HRTF filtering allows to present the audio in a more realistic way by incorporating the interaural time difference and level difference for a source at a particular location. It can also make the task more difficult because the signal-to-noise ratio in each ear under the HRTF condition is much lower than in the dichotic condition where each ear observes a source with quasi-infinite SNR. Since HRTF-filtered stimuli help bring the experiment conditions closer to the real auditory environment conditions, the observation that there is a significant performance difference between the 'dichotic' and the more realistic 'HRTF' condition is of interest. It is observed that the (more realistic) HRTF condition results in a higher accuracy, which is good news, in particular when targeting the use of AAD in real applications such as, e.g., hearing devices [5].

A question that then arises is *why* the HRTF condition results in a higher AAD accuracy, despite the fact that the per-ear SNR is lower than in the dichotic condition. This could be because the HRTF condition makes the task easier, e.g., because it is closer to realistic scenarios in which our auditory system usually operates, whereas the dichotic condition, being less realistic, could have a poorer cortical representation. On the other hand, it could also be the result of the task being more difficult in the HRTF condition, requiring additional effort from the subject, and hence resulting in larger cortical responses. To investigate which condition is more difficult, we conducted an additional behavioral experiment under the same set of stimulus conditions, the details of which are as follows.

Speech Recognition Threshold (SRT) experiment: An SRT experiment was conducted to investigate the effect of the different stimulus conditions used in the AAD experiment on speech recognition. Four subjects who were also part of the AAD experiment, participated in the SRT experiment. The experiment was essentially a 'speech in noise' test where the goal was to find the signal-to-noise ratio (SNR) at which the subject can repeat 50% of the words

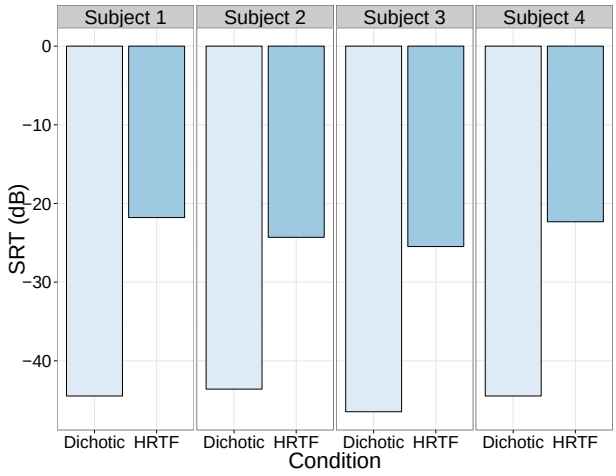

Fig. 4: Speech recognition thresholds per subject per condition. The SRT experiment uses sentences from the VU sentence database [27]. A short story from the AAD experiment is the masker.

in a sentence spoken in the presence of noise, which is referred to as the SRT. In this SRT experiment, the VU sentence database [27] was used to generate the target speech (this is a standardized database for SRT experiments). The noise in this case consisted of a competing speaker which was one of the stories presented during the AAD experiment. The subjects were asked to listen to and repeat the sentences presented in one ear, while ignoring the story being played in the other ear. The stimuli were administered in both 'dichotic' as well as 'HRTF' conditions, and the resulting SRTs were noted down. We adopted an adaptive procedure for measuring SRT, where the SNR was varied in steps of 2 dB from one trial to the next. If the subject repeated the sentence correctly during a trial, the SNR was lowered by 2 dB in the next trial. Similarly, if the subject couldn't repeat the sentence correctly, SNR was increased by 2 dB for the next trial. A total of 13 sentences were presented, and the SRT was calculated as the mean SNR from the last 6 trials. The subjects had to listen to their left and right ear alternately, and a total of 4 SRTs (one for each attended ear and each stimulus condition) were noted. Fig. 4 shows the average SRTs per subject. There was a clear indication that speech recognition was much easier under the dichotic condition, with 15-20 dB difference in SRT between the two conditions, consistently over all subjects. Thus, in our experiments, while both conditions were easy, the HRTF condition might have required a higher listening effort to focus attention to a speech stream in comparison to the less realistic dichotic condition.

From the results of the SRT experiment, and the AAD experiment, we hypothesized that, possibly, the 'dichotic' being an easier condition for speech recognition, might have a shorter processing time in the brain. To test our theory, we looked at the performance of AAD using different single-lag decoders with EEG signals shifted by 25 ms, 50 ms, and so on, until 250 ms. Here, single-lag decoders correspond to equation (1) with $N_l = 1$, after first shifting the EEG recordings back in time over a predefined time shift. If our hypothesis was true, over the range of these shifts, the AAD performance for the dichotic condition should have a peak/optimal performance for a shift smaller than that of the HRTF condition, indicating faster speech processing in the dichotic condition. However, as

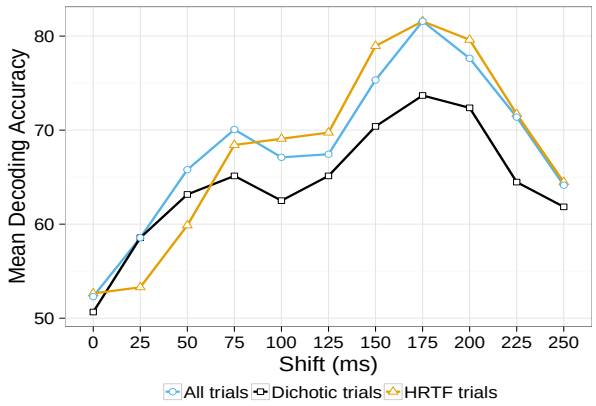

Fig. 5: AAD performance for single-lag decoders trained under different conditions (1-9 Hz).

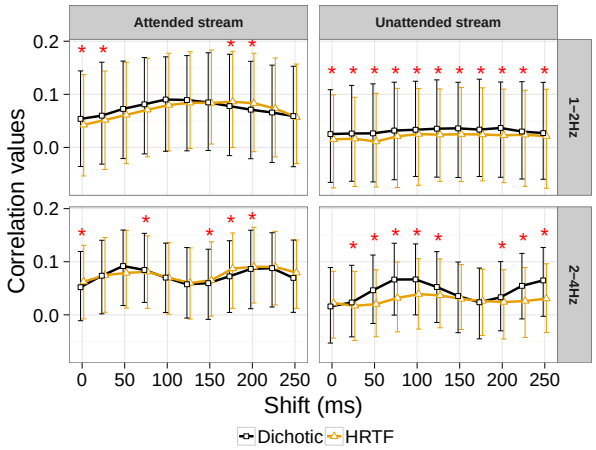

Fig. 6: Median correlations between original and reconstructed speech envelopes of attended and unattended streams, for 1-2 Hz and 2-4 Hz frequency bands.

can be seen from Fig. 5, contrary to our hypothesis, AAD performance for the two conditions, followed the same trend from a shift of 25 ms to 250 ms.

The AAD experiment showed that the HRTF condition results in better performance, while the SRT experiment revealed that HRTF-filtering, leads to poorer speech recognition. The results lead us to hypothesize that the better AAD performance in the HRTF condition could be attributed to a higher listening effort on the part of the subject leading to a better representation of the attended speech stream in the EEG recording, when faced with a challenging auditory environment for speech recognition. In order to get a clearer picture of how the two competing speech streams are cortically represented, we trained attended and unattended decoders to reconstruct the envelopes of the attended and unattended streams, for both conditions. Unattended decoders were designed the same way as the attended decoders except that they reconstructed the unattended speech envelopes, and thus in this case, $\mathbf{r}_{MS}$ in equation (2) represented the cross-correlation between the unattended envelope and the EEG data. We used single-lag decoders to look at the cortical representations at different time lags, and considered the correlation of the reconstructed envelope with the original speech envelope, for dichotic and HRTF condition.

In order to analyze the contribution of different frequency bands we investigated the correlations when decoders were trained on information from different frequency bands. We observed interesting effects within the 1-2 Hz and 2-4 Hz frequency bands. Wilcoxon's signed rank tests with Holm-Bonferroni (HB) correction were used to compare the differences in the correlation values between the reconstructed and original speech envelopes in the two conditions. The red stars in Fig. 6 indicate significant differences between the correlation values between the 2 conditions. We observed that during the early lags, the correlations of both attended and unattended streams are higher for the dichotic condition compared to the HRTF condition. Additionally, we also observed that in the later lags, where the overall attention detection performance also is seen to peak (see Fig. 5), the correlations of the

attended stream under the HRTF condition were significantly higher than that of the dichotic condition in both the 1-2 Hz and the 2-4 Hz frequency bands. For the unattended streams, during the later lags in the two frequency bands, and particularly in the 2-4 Hz band, the HRTF trials' correlations were significantly lower than those of dichotic trials. This showed that, at these lags, there was a poorer reconstruction of the unattended stream under the HRTF condition than under the dichotic condition. This leads us to believe that, during the later lags, in comparison to the dichotic condition, the cortical response to the unattended stream was suppressed to a greater extent, and the cortical response to the attended stream was strengthened, under the HRTF condition.

Thus, under the HRTF condition, we observed an increase of the attended stream as well as a suppression of the unattended stream at certain lags, which can both have an advantageous effect on the AAD performance. These findings could indicate differences in the way the human auditory system handles different acoustic situations. During the early lags, speech segregation is yet to happen. In the dichotic case, each ear has either only the attended or the unattended speech present at its periphery. If the earlier responses are thus mainly a peripheral representation of the sound, in dichotic condition, correlations could be expected to be higher for both the attended and the unattended streams during the early lags, since the peripheral streams in this case are 'clean'. If we consider the HRTF scenario, during the early lags, the sound at each ear is a mixture of the attended as well as the unattended streams. In this case, the periphery of each ear contains both streams and hence we would expect to see lower correlations with the individual streams, in comparison to the dichotic case. In Fig. 6, this effect can be seen in the early lags of both the attended and the unattended streams, where correlations values are significantly higher for the dichotic case than for the HRTF case. Furthermore, the dichotic condition, being a simpler listening scenario, possibly results in segregation of the two streams into auditory objects quite easily, and hence not producing a strong suppression of the unattended stream or a strong increase of the attended stream in the cortical activity. We believe that, in comparison to the dichotic condition, the better performance of the comparatively harder HRTF condition could be due to a combination of two effects: increase of its attended stream as well as suppression of its unattended stream during the later lags.

V. Conclusion

Auditory attention detection is an emerging research field that has the potential to support the signal processing in future hearing prostheses and possibly other BCIs. The focus of this paper is to understand the influence of head-related filtering and ear-specific decoding bias on the performance of an AAD algorithm, to better equip researchers to design their AAD experiments robustly and obtain reliable and representative performance estimates. The AAD experiments we have so far seen in the literature are different in their stimulus conditions and hence make it challenging to make comparisons between different studies.

We compared the performance of an AAD algorithm under two acoustic conditions, and found that the more realistic HRTF filtered stimuli result in better auditory attention detection compared to the simpler dichotic stimuli. A follow-up speech recognition threshold experiment under the same set of stimulus conditions revealed that the HRTF condition is a more difficult listening scenario for speech recognition. We have shown that, the attended

stream is easier and the unattended stream is more difficult to decode from the EEG recordings in the HRTF condition, which indicates that the cortical response to the attended speaker is higher and to the unattended speaker is more suppressed under this condition.

Another interesting observation in this study was the better AAD performance while using generic decoders trained on right-ear attended trials only, in comparison to those trained on left-ear attended trials. We believe what we observe here may be related to the right ear advantage. We also see that decoders that are trained on a single attended ear result in a higher AAD accuracy. This bias indicates that a balanced set of experiment conditions where the decoder is trained on both left ear trials and right ear trials per subject, is important to draw conclusions that are representative for real applications. Realistic experimental conditions are crucial to bringing this research one step closer to realizing real-time neurofeedback that supports hearing prostheses.

ACKNOWLEDGMENTS

The authors would like to thank Jonas Vanthornhout and Don Fleuren for their help in setting up and conducting the experiments. We would also like to thank all test subjects for their participation in the experiment.

REFERENCES

[1] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon *et al.*, "Mechanisms underlying selective neuronal tracking of attended speech at a Şcocktail partyŤ," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[2] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, p. bht355, 2014.

[3] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[4] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, 2015.

[5] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *Accepted for publication in IEEE Transactions on Biomedical Engineering*, 2016.

[6] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[7] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002.

[8] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13 367–13 372, 2001.

[9] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and hearing*, vol. 29, no. 2, pp. 139–157, 2008.

[10] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European journal of neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[11] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS-Biology*, vol. 10, no. 1, p. 175, 2012.

[12] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a Şcocktail partyŤ," *The Journal of Neuroscience*, vol. 30, no. 2, pp. 620–628, 2010.

[13] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[14] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a cocktail party," *Journal of neural engineering*, vol. 11, no. 4, p. 046015, 2014.

[15] A. J. Power, J. J. Foxe, E.-J. Forde, R. B. Reilly, and E. C. Lalor, "At what time is the cocktail party? a late locus of selective attention to natural speech," *European Journal of Neuroscience*, vol. 35, no. 9, pp. 1497–1503, 2012.

[16] S. Coren, "The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness: Norms for young adults," *Bulletin of the Psychonomic Society*, vol. 31, no. 1, pp. 1–3, 1993.

[17] T. Francart, A. Van Wieringen, and J. Wouters, "Apex 3: a multi-purpose test platform for auditory psychophysical experiments," *Journal of Neuroscience Methods*, vol. 172, no. 2, pp. 283–293, 2008.

[18] "Radioboeken voor kinderen," http://www.radioboeken.eu/kinderradioboeken.php?lang=NL, 2007, [Online; accessed: 30-March-2015].

[19] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[20] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand, "Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario." in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015.

[21] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 2016, accepted for publication.

[22] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[23] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The technology of binaural listening*. Springer, 2013, pp. 33–56.

[24] M. Hiscock and M. Kinsbourne, "Attention and the right-ear advantage: What is the connection?" *Brain and cognition*, vol. 76, no. 2, pp. 263–275, 2011.

[25] D. Kimura, "Functional asymmetry of the brain in dichotic listening," *Cortex*, vol. 3, no. 2, pp. 163–178, 1967.

[26] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[27] N. J. Versfeld, L. Daalder, J. M. Festen, and T. Houtgast, "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1671–1684, 2000.